Quantitative Understanding in Biology Short Course
Session 5
Power Calculations and Experimental Design

Jason Banfelder
Luce Skrabanek

April 8th, 2025

1 Introduction

In previous sections, we've seen over and over again that confidence intervals computed by statistical tests will be narrower in experiments that include more samples. In this section, we will use our knowledge of how CIs (and, equivalently, p-values) vary with n to plan experiments of an appropriate size.

Before we begin, it is important to recall that it is not valid to incrementally add samples to a study until you obtain a significant result. In other words, if you perform an experiment with a sample size of six and obtain a p-value of 0.07, you can't go back to the bench and add two more samples so you can rerun your statistics with $n = 8$. Since this probably isn't an argument you want to get into with your HOL, it is a good idea to carefully consider experimental design and sample size up front.

In this session, we'll define the precision of a CI as its half-width. In the case of a symmetric CI, the CI is written as

$$\text{CI} : \overline{x} \pm \text{precision} \tag{1}$$

Other texts may define precision differently, so if you look at other sources or use a computer program to run these calculations, make sure you know how this term is defined and adjust your interpretation accordingly. Some of the literature reasons in terms of "effect size". This is usually denoted with the symbol d, and is defined by the relation...

$$d = \frac{\text{precision}}{\text{SD}} \tag{2}$$

You can think of the effect size as a normalized precision. The SD and the precision have the same units of measure as the quantity being measured, so d is a dimensionless

quantity.

Statistical power calculations only give you estimates of the sample size that will allow you (with some likelihood) to conclusively observe a desired effect size, or one that is larger. Furthermore, when you use these methods, you'll need to estimate quantities like the standard deviation (SD) of the quantities that you'll measure. The bottom line is that most of what is presented in this section is approximate, so (1) we'll feel free to use approximations in our formulae, and (2) it is a good idea to be conservative when we provide our estimates.

2 Single Mean

From our previous lectures, we already know enough to estimate sample sizes for some special cases. Recall that the 95% CI for a univariate distribution with large n is given by...

$$95\%\text{ CI: } \overline{x} \pm 1.96 \cdot \text{SEM} \tag{3}$$

$$95\%\text{ CI: } \overline{x} \pm 1.96 \cdot \frac{\text{SD}}{\sqrt{n}} \tag{4}$$

If we rearrange and take $1.96 \cong 2$, then we can write...

$$n \cong 4 \cdot \left(\frac{\text{SD}}{\text{precision}} \right)^2 \tag{5}$$

This is a useful rule of thumb to have at your disposal for estimating how many measurements need to be taken to estimate the true mean to a desired precision.

Note that in order to use this formula, you'll need to estimate the SD of a population that you haven't taken samples from yet. You can usually get a rough idea of what this quantity will be by looking at previously obtained data. If you're not sure, be conservative and choose something on the high-side of what you expect.

The formula above only applies when n is sufficiently large to make the approximation that $t^* \cong 2$. So if you get $n = 4$ from the above formula, you should appreciate that you are likely to be underestimating n significantly.

This line of reasoning can be generalized by recalling that...

$$(1-\alpha)\,\text{CI: } \overline{x} \pm t^* \cdot \text{SEM} \tag{6}$$

...which implies...

$$n \cong \left(t^* \cdot \frac{\text{SD}}{\text{precision}} \right)^2 \tag{7}$$

Of course, t^* is a function of n (and α), so this equation has to be solved iteratively.

These sorts of calculations can be performed by a computer. While Prism doesn't perform power calculations like this explicitly, we'll show you in a moment how to use simulations to get the results you want. But it is useful to know that other software packages (such as R or StatMate) can do power calculations explicitly.

The above equations do not guarantee that if you perform n measurements, you'll obtain a CI with the desired half-width. In fact, if all of the assumptions in the analysis hold, you'll have a 50% chance of obtaining such a CI, or narrower. Put another way, **your power to obtain the desired precision will be 0.5**. The power of an experiment is an important quantity, and it is helpful to have an estimate of the power of an experiment before you perform it. Formally, the power of an experiment is one minus the probability of a type II error (assuming an effect of the specified size is actually present). Informally, power is the chance that you'll be able to measure an effect of a given size.

3 Difference Between Two Means

If you want to be able to determine the difference between the means of two groups of measurements to a certain desirable precision, the rule of thumb is...

$$n_{\text{each group}} \sim 8 \cdot \left(\frac{\text{SD}_{\text{each group}}}{\text{precision}} \right)^2 \tag{8}$$

There is an assumption that the SDs of the measurements from both groups are roughly the same. As before, the sample size given by this formula will give you a 50% chance of realizing your desired precision. You'll need significantly more samples for a 95% chance of hitting your target.

Example: In a series of knockdown experiments on MDCK cells, it was desired to confirm that preparations of the knockdown prevent the formation of functional tight junctions (TJ). This is assessed by measuring (among other things) transepithelial resistance (TER). Inspection of previous studies shows that the mean value of TER for wild-type cells that are known to form TJs is about $130\,\Omega\,\text{cm}^2$, and the standard deviation of TER measurements is about $30\,\Omega\,\text{cm}^2$. In this experiment, we are only interested in whether tight junctions form, not on the specific effects that a knockdown has on TER (perhaps via the regulation of TJs). We might say that variations of up to 35% in TER would still be indicative of TJ formation. The required precision for this experiment in therefore not particularly high: we just want a CI with a precision of roughly $\pm 45\,\Omega\,\text{cm}^2$. According to our rule of thumb, we'll need...

$$n_{\text{each group}} \sim 8 \cdot \left(\frac{30\,\Omega\,\text{cm}^2}{45\,\Omega\,\text{cm}^2} \right)^2 = 3.5 \tag{9}$$

This tells us that we'll need at least four samples per group. However, since the resultant n is small, we suspect this may be a significant underestimation.

It is pretty straightforward to model a single instance of this experiment in Prism. We begin by creating a Simulated Column Data analysis with two datasets, and four rows in each dataset. The mean of the first dataset (named WT) is 130, and mean of the second (named KO) is $130 \cdot (1 - 0.35) = 84.5$. Next, set up an unpaired t-test of this simulated data.

We've just set up a simulation for an experiment where we know there is a effect, since the means of the two columns of simulated data are not equal. However, we have very few data points, so even though the simulated effect is rather large, there is a reasonable chance that a t-test will not be able to conclude that the simulated data demonstrate a statistially significant difference bewteen the WT and KO data.

In fact, based on our back-of-the envelope calculation, we expect there to be a roughly 50/50 chance that the result of any given simulation will be statistically significant. You can convince youself of this by, in Prism, displaying the results from the unpaired t-test analysis, and asking Prism to redo the simulation with newly generated random data. This can be done by clicking the red icon that looks a die; each time you click, you'll see a new p-value computed from newly simulated data.

To more quantitatively assess the power of the simulated experimental design, we'd like to rerun the simulation a large number of times and observe how many cases yield a significant p-value. Fortunately, Prism can do this. With the result of the unpaired t-test displayed, add a new Monte Carlo analysis (found in the Simulated Data section). In the subsequent dialog, ask Prism to run 1,000 simulations, to tabulate p-values, and to classify a 'hit' as any simulation where $p < 0.05$ (for our current purposes, there's no need to tabulate results from individual simulations). When you're done, you will probably observe that Prism reports that about 44% of the simulated experiments are hits, meaning that our design has a power of a bit less than 0.5. If you go back to the specification of the simulation and change the number of samples to 5 each of the WT and KO samples, and then rerun the Monte Carlo simulation, you should see the power increase to about 56%.

Typically, we wouldn't bother investing in a experiment that had a 50% or so chance of finding an effect that we are interested in. We'd probably want a power of 80%, or even 95%. With a little bit of trial and error, you should be able to show that you'd need about 8 samples per group to have a power of 0.8, and around 13 samples per group to achieve a power of 0.95.

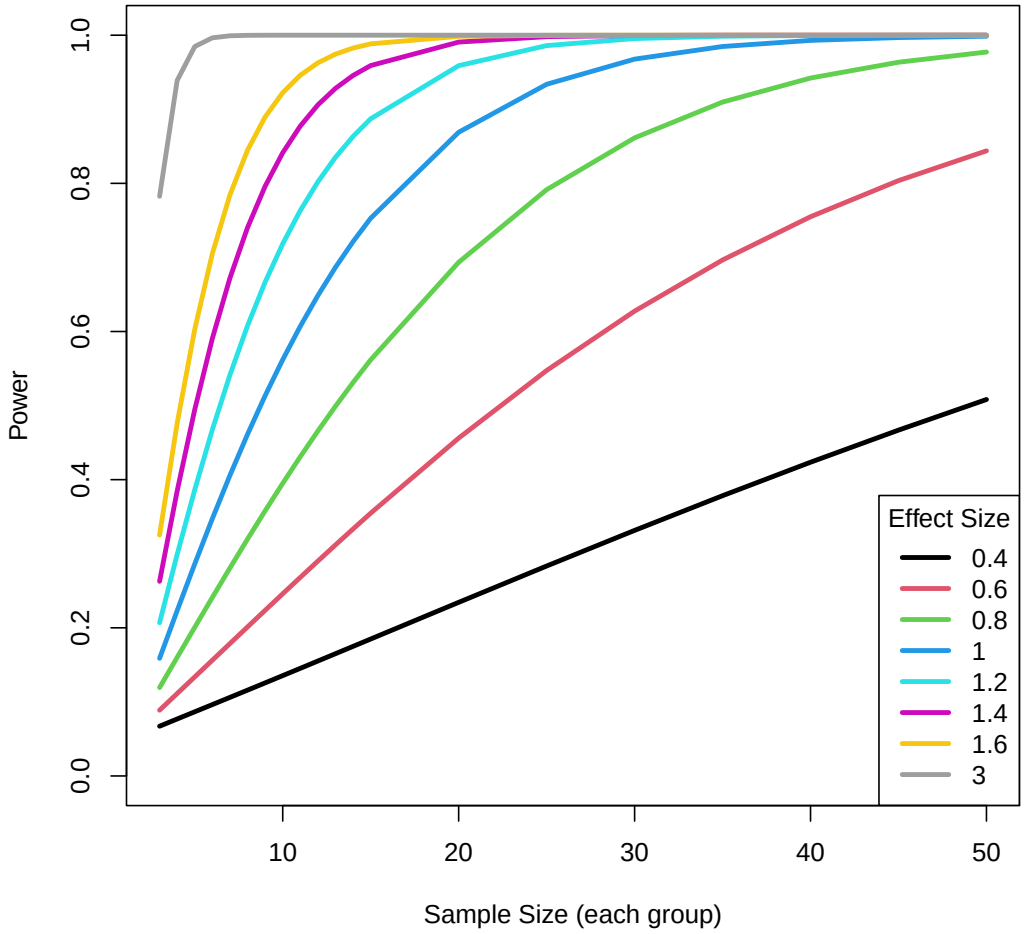
In the above example, we were hoping to detect relatively large effect sizes. If our experiment was looking not simply to determine if tight junctions were being formed, but rather to quantify potentially subtle effects of preparation methodology on TER, then we might say that we want to be able to resolve 10% changes in mean TER. Our precision would

then be $13\,\Omega\,\mathrm{cm}^2$, and our effect size would be $\frac{13}{30} = 0.43$. Our rule of thumb then tells us $n = 8(\frac{30}{13})^2 = 42.6$, so we estimate that we'd need 43 samples in each group to have a 50/50 chance of obtaining such a narrow CI.

This is confirmed by a simulation in Prism, and, with additional trial and error, we can show that we'd need about 140 samples in each group to achieve a 95% power. That's 280 samples in all, and assuming that you allow for some experimental problems, you likely need to plan (and budget) for 300 or so preparations.

If you thought that the above computed samples sizes were surprisingly high, you are not alone. Often when studies are planned (an all too rare event in the first place), the first power calculations along these lines can be quite depressing. Although we can use power calculations as above to compute a required n, budget, time and other constraints often put an upper bound on n. What is usually needed in practice is a more holistic view of the interplay and tradeoffs among power, effect size, and n that will aid in the selection of a pragmatic experimental plan.

Preparation of plots can be very helpful in this regard. This plot shows the power of experimental designs based on the canonical two-tailed unpaired t-test with equal SDs in each group, and with $\alpha = 0.05$.



Plots like these typically are the most useful for planning experiments. All curves will have the same basic sigmoidal shapes because there is always zero power as the sample size approaches zero, and power can be made arbitrarily high by increasing sample size to something very large.

If your power is very low, you may be better off not doing the experiment at all (this is always an option). Similarly, if your experimental plan puts you on the upper flat part of these curves, you might consider reducing your sample size a bit.

To review: In reality, the decision to include a certain number of samples in an experiment is driven not by a single power calculation, but by understanding the tradeoffs among

power, sample size, and precision. The precision you need (or want) is something that should be guided by your scientific judgment and understanding of the underlying biology of your system.

All of the results hinge on having a reasonable estimate of the variation of your data (this appears as the SD in these analyses). Recall that in many biological studies, variation can come from both measurement error and biological diversity. You can do something about measurement error by being more careful at the bench, or by switching to more precise methods, but realize that a good deal of intrinsic biological variation is typically unavoidable.

4 Non-Equal Sample Sizes

As mentioned above, the decision to include a certain number of samples in an experiment is usually driven in part by budget and time constraints. In some cases, the constraints on sample size may be hard limits if you only have access to a fixed number of consenting patients with a rare disease or a limited number of surgical tissue specimens. In many such cases, the hard constraint is imposed on the number of samples in one group only. You can still gain some statistical power by increasing the number of samples in the other group (typically the control group), but there are limits to this.

You'll always need the fewest total samples when sample sizes are equal, but you can use unequal sample sizes if you need to. For example, we need roughly 64 samples in each group to have an 80% chance of measuring an effect that is half the size of the SD of the data we are collecting. However, if we have access to only 48 experimental samples, we can still achieve 80% power by using 95 samples in the control group. The total number of samples, $48+95 = 143$ in this case, is more than the 128 needed in a balanced design. This could also be a useful trick to employ if creating experimental samples is substantially more expensive or time-consuming than control samples, even if the number of experimental samples is not strictly limited.

There are, however, limits to how far you can take this. To continue the example above, if you only have access to 30 experimental samples, you simply cannot measure an effect of size $d = 0.5$ with a power of 80%.

As above, plots can be prepared to gain insight into the tradeoffs that are at work under these circumstances.

5 Homework

Prepare a Prism workbook that confirms the number of samples needed for each of the cases in Section 4 above.

-or-

Describe an experiment that you or a colleague recently performed (or, even better, are about to perform) in the lab. What is the approximate SD of the measurements you make? What is the size of the biological effect that you want to find? Perform the relevant power analysis, and indicate where you are on the figure from section 3. In light of the result, are you happy with the design, or would you alter it?

